Organic Chemistry Predicting Outcomes

Zackary Scalyer and Katie E. Amaral, Ph.D.

Penn State Berks | P.O. Box 7009 Tulpehocken Rd. Reading, PA 19610

zes5027@psu.edu, kea12@psu.edu

Introduction

Modeling outcomes in Organic Chemistry is a means to understanding student success; however, there is limited literature on this topic. As an attempt to investigate academic performance, 7,894 students who had taken Organic Chemistry at a large state university with many campuses within a 3 year span were analyzed. General Chemistry I, General Chemistry II, cumulative grade point average, and gender were found to be significant indications for Organic Chemistry performance. Predictions were made and compared using multiple linear regressions and generalized linear regression analyses. With every model, there is some degree of uncertainty and assumptions that must be made; therefore, comparing and contrasting two different approaches for modeling student success may provide further insight into the Organic Chemistry classroom.

Conclusions

By re-classing the organic chemistry average GPA to a binomial outcome allows for a generalized approach to regression analysis. Logistic regression provides easily interpretable predictions and confidence intervals compared to the liner model. However, both methods have proportions of the residual distribution outside of the 95% confidence interval. The relationship between this population must be investigated with the possibility of multivariate regression.

Prior Research

Literature Cited

1. Rauschenberger, M.M. and Sweeder, R.D. "Gender Performance Differences in Biochemistry" *Biochem. and Molec. Biol Ed*, **2010**, *38*, 380-384.

2. Szu, E., Nandagopal, K., et.al. "Understanding Academic Performance in Organic Chemistry" *J.Chem. Ed.* **2011**, *88*, 1238-1242.

3. Field, A. P., Miles, J., & Field, Z. (2012). Discovering statistics using R. Thousand Oaks, Calif, London: Sage.

Generalized Linear Model

CHEM 210 Binomial outcome	Coefficients	Std. Error	Z-Value	P(Z-Value)
(Intercept)	-1.66	0.0655	-25.33	<2e-16
CHEM 112 average grade	0.343	0.0215	16.00	<2e-16
CHEM 110 average grade	0.068	0.0246	2.78	5.52e-3
Last known cumulative GPA	0.102	0.0062	16.31	<2e-16
Gender, Male	0.236	0.0334	7.07	1.53e-12
AIC:	7730.7			
Hosmer & Lemeshow's R-sqr:	0.173			
Chi-squared :	1619 on 4 DF			
P-Value:	<2.2e-16			
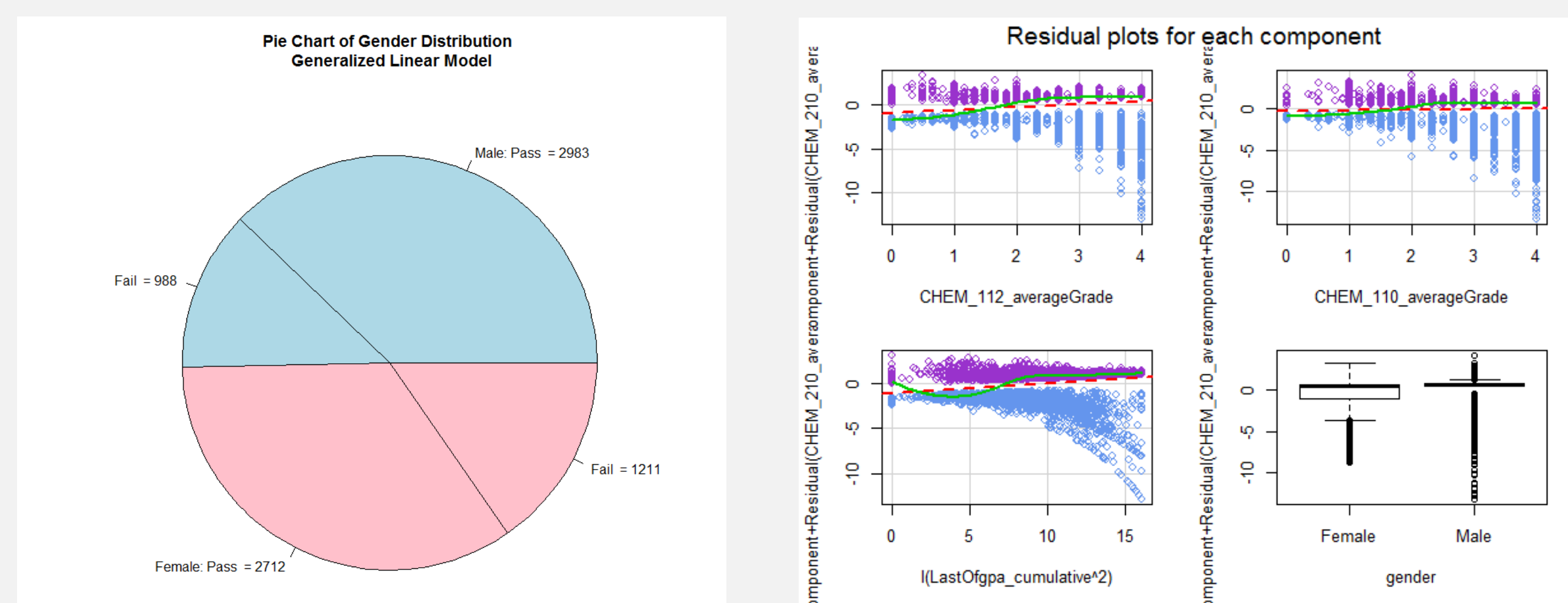
	Odds Ratio	2.50%	97.50%
(Intercept)	0.190	0.167	0.216
CHEM 112 average grade	1.41	1.352	1.47
CHEM 110 average grade	1.07	1.02	1.123
Last known cumulative gpa	1.11	1.094	1.12
Gender, Male	1.267	1.187	1.353

Multiple Linear Regression Model

CHEM 210 Average GPA	Coefficients	Std. Error	T-Value	P(T-Value)
(Intercept)	-0.321	0.0449	-7.138	1.03e-12
CHEM 112 average grade	0.377	0.0156	24.140	<2e-16
CHEM 110 average grade	0.106	0.0180	5.906	3.65e-9
Last known cumulative GPA	0.108	0.0045	23.961	<2e-16
Gender, Male	0.233	0.0230	10.124	<2e-16
Multiple R-squared:	0.3492			
Adjusted R-squared:	0.3488			
F-statistic:	1058 on 4 and 7889 DF			
P-Value:	<2.2e-16			
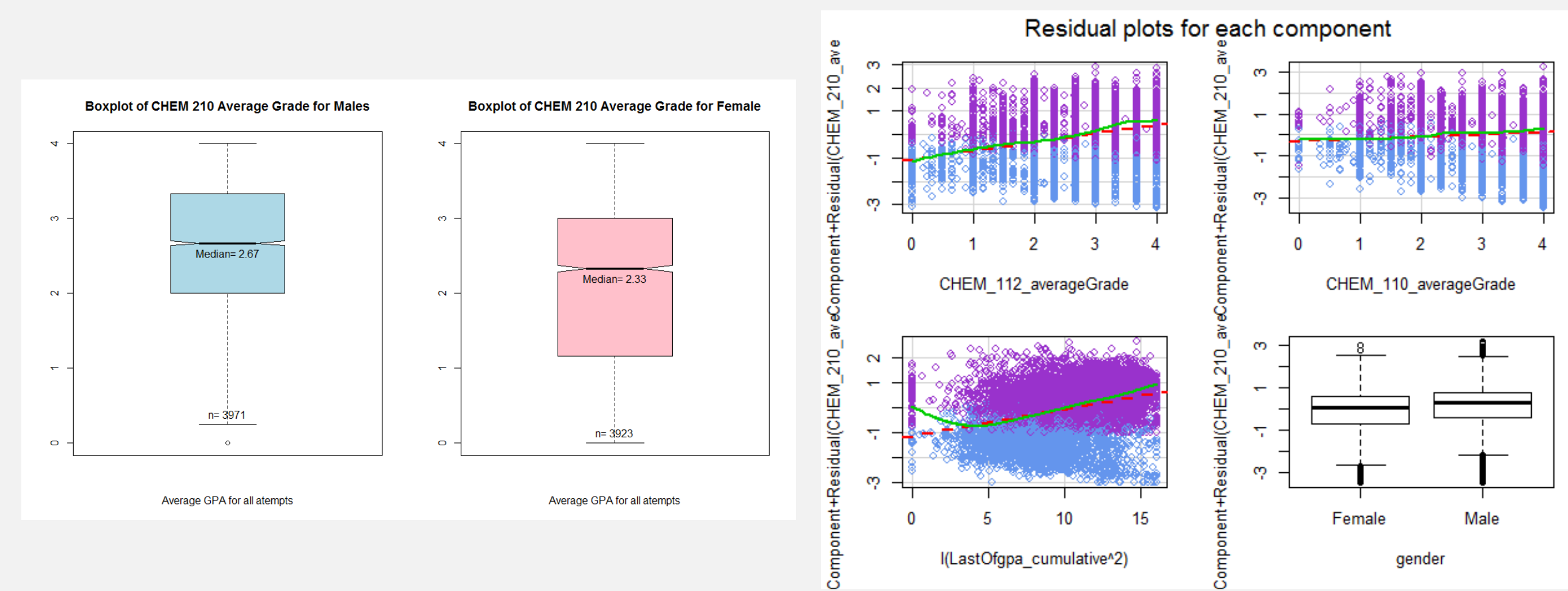
	Coefficients	2.50%	97.50%
(Intercept)	0.190	0.167	0.216
CHEM 112 average grade	1.41	1.352	1.47
CHEM 110 average grade	1.07	1.02	1.123
Last known cumulative gpa	1.11	1.094	1.12
Gender, Male	1.267	1.187	1.353

Logistic regression

i) Linearity: Residual plots
ii) Independence of errors: Residual plots
iii) Multicollinearity:
   - VFI for all variables is < 10
   - Average VIF = 1.4 could imply bias in the model

Outliers and influential case assessment
1. 186 cases outside ± 2 standard residuals
   - 186/7894 → with α = 0.05, 2.4% < 5% of cases
2. 13 of 186 cases outside ± 2.5 standard residuals
   - 13/7894 → 0.001 is < 1%
3. 63 of 186 out side ± 3 × expected covariance ratio
   - Possible undue influence over the model
4. 0 of 186 cases have cooks distance > 1
   - no cases have undue influence

Linear regression

i) Variable types: Predictors must be quantitative, or categorical with 2 levels. Outcome must be quantitative, continuous, and unbounded.
ii) Non-zero variance: Residual plots
iii) Multicollinearity: - VFI for all variables is < 10;
   - Average VIF = 1.6 could imply bias in the model
i) Homoscedasticity: - NCV test; p < 0.05
ii) Independent errors: Residual plots
iii) Normally distributed errors: out liar test; p = 0.003
iv) Independence: D-W Statistic = 1.56; p = 0.09
v) Linearity: Residual plots

Outliers and influential case assessment
1. 430 cases outside ± 2 standard residuals
   - 430/7894 → with α = 0.05, 5.4% > 5% of cases
2. 150 of 430 cases outside ± 2.5 standard residuals
   - 150/7894 → 0.0190 is > 1%
   - Possible violation of normally distributed error
3. 205 of 430 out side ± 3 × expected covariance ratio
   - Possible undue influence over the model
4. 0 of 430 cases have cooks distance > 1
   - no cases have undue influence

Predictions

Male with D Average	
CHEM 112 average GPA	1.000
CHEM 110 average GPA	1.000
Cumulative GPA	1.000
Gender	Male
Multiple Linear Regression	
0.504 (-1.47 – 2.48)	Grade Point Average
Generalized Linear Regression	
18.1% (15.5% - 20.8%)	Chance of Passing

Male with C Average	
CHEM 112 average GPA	2.000
CHEM 110 average GPA	2.000
Cumulative GPA	2.000
Gender	Male
Multiple Linear Regression	
1.31 (-0.661 – 3.28)	Grade Point Average
Generalized Linear Regression	
42.3% (39.5% - 45.1%)	Chance of Passing

Male with B Average	
CHEM 112 average GPA	3.000
CHEM 110 average GPA	3.000
Cumulative GPA	3.000
Gender	Male
Multiple Linear Regression	
2.33 (0.362 – 4.31)	Grade Point Average
Generalized Linear Regression	
76.6% (75.1% - 78.1%)	Chance of Passing

Male with A Average	
CHEM 112 average GPA	4.000
CHEM 110 average GPA	4.000
Cumulative GPA	4.000
Gender	Male
Multiple Linear Regression	
3.57 (1.60 – 5.56)	Grade Point Average
Generalized Linear Regression	
96.8% (96.1% - 97.4%)	Chance of Passing

Female with D Average	
CHEM 112 average GPA	1.000
CHEM 110 average GPA	1.000
Cumulative GPA	1.000
Gender	Female
Multiple Linear Regression	
0.271 (-1.70 – 2.44)	Grade Point Average
Generalized Linear Regression	
12.6% (10.4% - 14.7%)	Chance of Passing

Female with C Average	
CHEM 112 average GPA	2.000
CHEM 110 average GPA	2.000
Cumulative GPA	2.000
Gender	Female
Multiple Linear Regression	
1.08 (-0.89 – 3.05)	Grade Point Average
Generalized Linear Regression	
33.4% (30.6% - 36.1%)	Chance of Passing

Female with B Average	
CHEM 112 average GPA	2.000
CHEM 110 average GPA	2.000
Cumulative GPA	2.000
Gender	Female
Multiple Linear Regression	
2.10 (0.128 – 4.07)	Grade Point Average
Generalized Linear Regression	
68.8% (66.9% - 70.6%)	Chance of Passing

Female with A Average	
CHEM 112 average GPA	4.000
CHEM 110 average GPA	4.000
Cumulative GPA	4.000
Gender	Female
Multiple Linear Regression	
3.34 (1.37 – 5.31)	Grade Point Average
Generalized Linear Regression	
94.6% (93.8% - 95.5%)	Chance of Passing